Perceived Credibility in the Evaluation of Online Music Recommender Systems

Li Lu
University of Missouri-Columbia
School of Information Science and Learning Technologies
571-299-8529
llnvd@mail.mizzou.edu

Yunhui Lu
University of Missouri-Columbia
School of Information Science and Learning Technologies
573-823-9144
Ly55f@mail.mizzou.edu

Xiuzhenn Huang
University of Missouri-Columbia
School of Information Science and Learning Technologies
573-529-9374
xh8r4@mail.mizzou.edu

ABSTRACT
Credibility as a source characteristic has been found to be highly influential in human advice-seeking. Credibility has also been found to matter when computers gives advice or provide instructions to users. This study investigates the role of perceived credibility as well as other theoretically important variables in the evaluation of online music recommender systems. Think-aloud lab interviews and small-scale surveys are conducted to obtain users' feedback on their perceptions of the credibility of 2 music recommender systems, Pandora and Last fm. Our finding indicates that dimensions of credibility, expertise and trustworthiness, are potentially important predictors of users' attitude and behavioral intentions toward music recommender systems. Besides, transparency, required efforts of using the system, interface and interaction adequacy are found to be important cues of system credibility. Implications for human computer interaction and design of recommender systems are also discussed.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—Information filtering; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—Collaborative computing; H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedial—Navigation; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces—Evaluation/methodology

General Terms
Measurement, design, reliability, human factors, evaluation, design

Keywords
Credibility, expertise, trustworthiness, recommender system

1. INTRODUCTION
The development of intelligent search and recommender systems can provide people with tailored information and content that is pertinent to their preferences and use contexts. Besides its critical importance for reducing information overload, comparing to general information, tailored information increases the potential for attitude and behavior change (Beniger, 1987; Dijkstra, Librand, & Timminga, 1998; Jimison, Street, & Gold, 1997; Nowak, Shamp, Hollander, Cameron, Schumann, & Thorson, 1999; Strecher, 1999; Strecher, Kreuter, Den Boer, Kobrin, Hospers, & Skinner, 1994).

Recommender systems are often regarded as one of the most promising applications for electronic commerce (Spiekermann & Paraschiv, 2002). Music recommender systems are a type of web technologies that proactively suggest music items of interest to users based on their objective behaviors or their explicitly stated preferences (Pu & Chen, 2010). They recommend music to users after a preference-elicitation process, such as rating different music, choosing specific types of music, or by simply specifying the current mood of the user (like in Musicovery.com).

Burke pointed out that recommender systems differ from other information-retrieval systems and search engines in that they not only match and return every entry that matches the query but also emphasize relevance and usefulness, and often individualize the information they present (Burke, 2002). Researches also suggested that making information relevant to individuals increases their attention and arousal, which can ultimately lead to increased attitude and behavior change (Stretcher, 1999).

Perceived qualities of a recommender system can influence users' attitudes as well as behavioral intentions. A satisfactory recommender system is supposed to make the users: 1) use the system as often as possible and 2) continue to use it in the future, 3) purchase items via the system, and 4) recommend this system to their friends. User experience has been the center of recommender system research. For example, in the study of Swearingen, the quality of recommendations and usability of three book RS (Amazon.com, RatingZone & Sleeper) and three movie RS (Amazon.com, MovieCritic, Reel.com) were examined. They found that, from a user's perspective, an effective recommender system inspires trust in the system; has system logic that is at least somewhat transparent; points users towards new, not-yet experienced items; provides details about recommended items, including pictures and community ratings; and finally, provides ways to refine recommendations by including or excluding particular genres. Users expressed willingness to provide more inputs to the system in return for more effective recommendations (Swearingen& Sinha, 2001).

ResQue Model, which was developed by Pu and Chen in 2010, presented a comprehensive model to evaluate the perceived qualities of recommender system (Pu & Chen, 2010). This model consists of 13 constructs and aims to assess the perceived qualities of recommenders such as their usability, perceived usefulness, interface and interaction qualities, users' satisfaction of the systems, and the influence of these qualities on users' behavioral intentions, including their intention to purchase the products recommended to them, return to the system in the future, and tell their friend about the system.

The review of existing studies suggests that perceived system credibility is an important dimension in users' evaluation of a

recommender system. Classical communication research suggested credible source have the ability to change opinions, attitudes, and behaviors, to motivate and persuade. In contrast, when credibility is low, the potential to influence also is low (Hovland &Weiss, 1951). As defined by B. J. Fogg, credibility is a perceived quality that has two dimensions: trustworthiness and expertise. The trustworthiness dimension of credibility captures the perceived goodness or morality of the source. Computing technology that is viewed as trust-worthy (trustful, fair, and unbiased) will have increased power of persuasion. The second dimension of credibility is expertise—the perceived knowledge, skill, and experience of the source. Computing technology that is viewed as incorporating expertise (knowledge, experience, and competence) will have increased power of persuasion (Fogg, 2003).

However, as pointed out by J. B. Fogg, trust and credibility are often used imprecisely and inconsistently (Fogg, 1999). Trust and credibility are not the same concept. Trust indicates a positive belief about the perceived reliability of something. Simply put, credibility means believability, while trust means dependability. So, it is important to distinguish these two different concepts and see whether perceived credibility of a music recommender system has an impact on users' evaluation of the sites.

Transparency has been studied quite extensively in the context of recommender systems (Herlocker, Konstan & Riedl, 2000). Bilgic and Mooney argue that a system's ability to make its reasoning transparent can contribute significantly to user's acceptance of the system's suggestions (Bilgic & Mooney, 2005). Process transparency is believed to increase the perceived value and overall acceptance of recommender systems (Kwak, 2001). Consequently, transparency appears to be an important factor in determining whether a user will like to accept the recommendations made by the system.

Effort is also a critical concept in the context of decision aids. Cognitive effort is typically seen as a cost users seek to avoid or at least reduce. Asking the user to provide information to the system so that the system can provide personified recommendation is typically regarded as an undesirable burden on the user. Although some suggest that individuals do not always try to reduce effort, but instead attempting to make a reasonable amount of effort so that that can achieve a more desirable result. We argue that, in the specific case of music recommender system, users are often more relaxed and not faced with an important purchasing decision to make comparing to other commercial products recommender system using context, so they are not willing to put a considerable amount of cognitive effort in the preference elicitation process, even it could yield more accurate result. According to the Norman's theory of emotional design, people are more tolerant when they feel relaxed and pleased, and it is more often the case when they are listening to music. Besides, a lot of people would listen to personalized online music radio when they in a moving status. So it's unlikely that the users would like to put in much effort to find what they want to listen. Although effort seems to lead to better evaluations of the outcome, it is negatively correlated to satisfaction with the process (Bechwati & Xia, 2003). So, it's necessary to look at the effort that the users are willing to put in a music recommender system, and how it would affect their attitudes and behaviors.

2. RESEARCH MODEL AND HYPOTHESES

Based on this theory and research illustrating the persuasive impact of technology (Fogg, 2003), it is possible that the structure of the system and the preference elicitation task can substantially influence users' attitude towards the recommendation provided by a recommender system, their behavioral intentions concerning those recommendations, especially their evaluations of the recommendation's matching with their interests or needs. We propose four possible factors that can influence users' perceptions of how well the recommendation matches their preferences: (1) transparency, (2) perceived qualities, (3) perceived credibility, and (4) effort.

In order to find out what are the essential qualities of an effective and satisfying recommender system and the essential determinants that motivate users to adopt this technology, we propose the following evaluating model for music recommender system shown in Figure 1.

Hypotheses

We propose the following four hypotheses:

H1. Perceived interaction adequacy and perceived interface adequacy are positively related to expertise.

H2. Perceived interaction adequacy and perceived interface adequacy are negatively related to effort.

H3. Transparency, expertise and trustworthiness are positively correlated to attitude and behavior.

H4. Effort is negatively correlated to attitude and behavior.

Two music recommender systems, Pandora and Last fm, will be evaluated in this study. Pandora Radio is an automated music recommender system. Its music data is based on a music project called the Music Genome Project. Users enter a song or artist that they like, and the system responds by playing selections that are similar. Users provide feedback to system about whether they like (*thrums-up*) or dislike (*thrums down*) individual songs or a series of songs, which they call "station. Besides online, Pandora can be used on mobile devices and also installed in some vehicles[1].

Last.fm is a more comprehensive music website. But it also has a personal music radio function which is similar to Pandora. For our current study, we only compare this part of the whole Last fm to Pandora systems. User starts using the radio by typing in their favorite artists in their library. System then generates a station based on the users' input.

For both systems, recommendations are calculated using a collaborative filtering algorithm so users can browse and hear previews of a list of artists not listed on their own profile but which appear on those of others with similar musical tastes. Both systems allow users to manually recommend specific artists, songs or albums to other users on their friends list (Pandora does it with users' Facebook account) or groups they belong to.

3. METHODOLOGY

We designed and conducted a pilot study to test the importance of credibility on users' attitudes towards the two music recommender systems. The study included ten music lovers, five female and five male. The range of age is 20 to over 50. Six participants are current doctoral students with an earned master's

[1] http://www.pandora.com/auto

degree. Most of the participants rate their own computer and Internet skills as strong (4 or 5 on a Likert scale). According to the demographic survey, all participants listen to music very often, either every day or several times per week. In order to avoid any preceding bias, we recruited only new users of Pandora and Last.fm.

The study consisted of two sessions. During the first one, participants' task was to create a new account with both systems and to interact with each one for about twenty minutes. They were asked to tell the systems their music preferences and see whether the recommendations provided by the systems met their interests. They were encouraged to explore available features of the systems. A think-aloud protocol was employed for data collection. Participants were encouraged to share with the interviewer what they like and what they don't. But, during the session, they were required to finish tasks by themselves. The interviewer did not give any hints or suggestions; instead, interviewer observed participants' interaction with the systems.

The second session took place one week after the first one. During that week between the two sessions, participants were asked to use both systems at home for at least 30 minutes. With their permission, we also logged in their accounts and kept track of their usage. In the second session of study, participants spent around five minutes interacting with Pandora, and filled out a questionnaire for Pandora. Then they spent five minutes on Last. Fm and filled out a questionnaire for it. At the end of the session, participants answered interviewer's questions about their preference of system and behavioral intention. Qualitative data about decision reasoning were collected.

The survey instrument to measure the dimensions of credibility and other elements of recommender usability was developed based on the credibility model proposed by Fogg (2003) as well as Yoo and Gretzel's (2006) instrument for measuring the credibility of recommender systems, Beyah alt.'s (2003) instrument to study the use of recommendation systems, and Pu and Chen's (2010) ResQue framework for evaluating recommender systems. We defined "perceived quality of recommended items" as "expertise" because these two concepts have similar constructs in the above mentioned literature. "Trustworthiness" was defined as the combination of reliability and intentions (Delgodo-Ballester, 2004). "Reliability" was conceptualized as consistency in performance of the system, and intentions as users' perceptions of system's purpose and motives. A questionnaire including 71 questions was designed to measure the systems' interaction adequacy, interface adequacy, expertise, transparency, effort, trustworthiness, overall attitude toward the system, and behavioral intention after using the system. All questions were measured on a 5-point Likert scale ranging from 1-strongly disagree to 5-strongly agree.

Simple statistical analysis was conducted to see whether there is correlation between proposed variables and what the directions of the correlation are. Qualitative data generated from the interviews will be the focus of analysis.

4. RESULTS AND DISCUSSION

The goal of our analysis was to find out whether credibility played an important role in affecting user's attitude and behavioral intention as comparing to other elements such as interface, interaction, efforts required in preference-elicitation process, and transparency, and how credibility would affect user's perception of the system. To answer these questions, we analyzed the data we gathered in the study: answers to semi-structured interview questions, survey about each variables in the model, self-report comments during test and observation made by facilitator. Because of the length limitation, the qualitative findings about each variable will not be presented here. Detailed discussion about credibility aspects will be discussed subsequently. We also offer some design suggestions for music recommendation systems and implications for human computer interaction at the end of this part.

4.1 Overall finding

Overall, participants have a more favorable attitude towards Last fm. They are more likely to use Last fm in the future other than Pandora (see Figure 2). The overall finding from the quantitative statistic also showed consistent results.

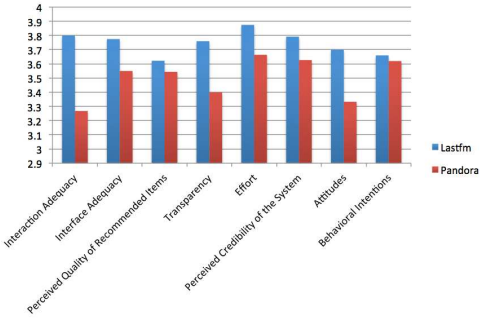

Figure 2 Overall finding for Pandora and Last fm

4.2 Correlation analysis of quantitative data

Spearman correlations were conducted to examine the relationships between dimensions of credibility and participants' attitude and behavioral intention toward both systems (see Figures 9 and 10). Perceived interaction adequacy and perceived interface adequacy are not significantly related to expertise, so H1 is not supported. Perceived interface adequacy is negatively correlated to effort (the scale for effort is reversed, so the coefficients are positive) for both Pandora and Last fm. Perceived interaction adequacy is significantly correlated with effort in Pandora, but it's not significant for Lat fm. So H2 is partially supported. Besides, the effect of effort on the attitude is also different for Pandora and Last fm. Effort is significantly correlated to attitude for Pandora but not for Last fm. We assume that people prefer Last fm not because it's interaction adequacy or because it's easy to use, but for other reasons, (we will discuss them in the limitation part). Thus, H4 is partially supported.

Participants' perceptions of transparency of both systems are different either. Transparency is significantly correlated to attitude for Pandora but not for Last fm. On contrary, expertise and trustworthiness are significantly related to users' overall attitude towards both systems, which implies that dimensions of credibility are potentially important predictors of users' attitude and behavioral intentions towards both systems. So H3 is also partially supported.

4.3 Discussion

The overall finding is consistent with Swearingen's finding about the characteristics that an effective recommender system should have (Swearingen, 2001). Last fm provides more diverse recommendations, has less advertisement, includes more social elements in the sites and is perceived by users as being more neutral and has better intention. It's not surprising that

participants have a more favorable attitude towards it. Swearingen suggested that a good recommender system inspires trust in the system; and have the system logic transparent to users; and points users towards new, not-yet-experienced items; including pictures and community ratings, provides ways to refine recommendations by including or excluding particular genres (Swearingen, 2001).

Our finding supported the existing literature on the importance of provide trust-inspiring interface to convince users of the systems' recommendations (Pu & Chen, 2007). The perceived credibility of systems seems to be even more important than interface or interaction adequacy in determining users' attitude toward the system, though these elements are closely correlated. A trust-inspiring system is perceived by users as more capable and efficient in assisting them to find what they would like to listen or purchase, and they are more like to use the system in the future.

Our qualitative findings about the interface and interaction adequacy and credibility are also consistent with Fogg's study on how users evaluate the credibility of web sites (Fogg et al., 2003b). Fogg's large scale survey yields the result that the design look of the site, information structure and information focus are the most mentioned criterion when people comments about web credibility, followed by underlying motive(good intention). More than half of the participants of our study mentioned that they prefer a system because the system interface is more appealing to them than the other one, though our correlation results contradictedFogg'sfinding.

Measures	1	2	3	4	5	6	7	8
1 Interaction	1	.196	.307	.432	.274	.403	.164	.300
2 Interface			.557	.629	.668*	.544	.492	.064
3 Expertise				.774**	.899**	.753*	.884**	.561
4 Transparency					.923**	.865**	.728*	.474
5 Effort						.841**	.782**	.396
6 Trustworthiness							.689*	.652*
7 Attitude								.720*
8 Behavioral intention								1

Figure 3 Correlation Table of Pandora

Measures	1	2	3	4	5	6	7	8
1 Interaction	1	.599	.506	.353	.864**	.412	.449	.491
2 Interface			.433	.249	.733*	.502	.557	.718*
3 Expertise				.508	.402	.670*	.784**	.704
4 Transparency					.563	.316	.530	.602
5 Effort						.529	.565	.613
6 Trustworthiness							.927**	.633*
7 Attitude								.693*
8 Behavioral intention								1

Figure 4 Correlation Table of Last fm

5. Implications for HCI and Design

Based on our findings, few constructive guidelines have been established for designing a recommender system. First, function buttons should be more visible for users as they would not like to contribute too much effort for a specific task. For instance, Pandora has many good but hidden features. Users do not realize they are available at all. Designers cannot suppose users would spend a long time digging through to find them. They should make them stand out and easy for users to see.

Second, designers can implement universal usability in recommender systems. Shneiderman's (2003) multi-player interface design would be applicable. The design should give novice users a limited set of features while provide expert users with advanced features. Some users, especially older users only need basic functions of the system; a complicated system with fancy features can only confuse and frustrate them. Designers can help accommodate their struggle.

Third, give users more control. This is also consistent with Norman's behavioral level of a good design (Norman, 2004). Users like using the system when they feel everything is under their own control. However, limitations of systems such as limited numbers of songs can be skipped, users unable to bookmark specific songs and no control of volume on Pandora make users feel the system rather than themselves is making decisions during the process. Therefore, if possible, the system should support users' control.

Fourth, make ads on the system less dominant. Commercial and ads have great influence on users' attitudes to the system. Too much ads can overwhelm users and even worse, can make the system perceived as less neutral and biased. According to users' feedback, static ads are less annoying and users feel more tolerant when they are given the information of how long the commercial will last.

Last, the logic of how the system works is important to users. The user feedback on transparency focused on user does not understand why a recommendation is provided to him. This has negative influence on users' trust on the system. Designers can provide a brief explanation about why a product is recommended to users so that they can better understand the logic of the system and better make use of it

6. Implications for HCI and Design

This study is a pilot study of the credibility issue in the evaluation of music recommender system. Music recommender system has its own limitations, and so the result is not generalizable to other types of recommender systems. Due to time and manpower limitations, we can only recruit ten participants to test the two systems. It is not a sufficiently large number for a regression test about the relationship between credibility and attitude and behavior changes. We can only predicate a trend about the relationship between them. Also, because there is only one week for each user to explore the two musical systems, it is very likely that the users did not get enough information and experience with the systems and their feedback might not be very accurate.

Given sufficient time, we would recruit more participants with diverse background to participate the study. Also, we would like to expand the time between the first and second interview time so that the participants have enough time to explore the two systems. Large scale survey would be helpful in determine the exact cause effect relation between variables and their specific explanation power.

Based on previous researches and this study, we can conclude that increasing a site's credibility can affect a user's attitude and behavior. With the above mentioned constraints and limitations, more thorough and nuanced researches need to be done before we can make more specific and meaningful suggestions to the design of music recommender systems. This study only serves to provide an initial set of results that future research on music recommender systems or other types of recommenders can refine or refute. In order to make persuasive technology framework applicable to other commercial recommender systems, more general model and corresponding researches on other types of recommender systems are also need to be done.

7. ACKNOWLEDGMENTS

Authors would like to thank the instructor of the human computer interaction course, Dr. Joi Moore.

8. REFERENCES

[1] Bechwati, N.N., and Xia, L.(2003) Do computers sweat? The impact of perceived effort of online decision aids on consumers' satisfaction with the decision process. *Journal of Consumer Psychology, 13*, 1–2 (2003), 139–148.

[2] Beniger, J. R. (1987). Personalization of mass media and the growth of pseudo-community. *Communication Research, 14*(3), 352–371.

[3] Bilgic, M., & Mooney, R.J. Explaining recommendations: Satisfaction vs. promotion. In M. van Setten, S. McNee, and J. Konstan (eds.),New York, ACM Press, 2005 (www.grouplens.org/ beyong2005/bp2005.pdf).

[4] Burke, R. Hybrid recommender systems: Survey and experiments. *User Modeling and User*-Adapted Interaction, 12, 4, 331–370.

[5] Dijkstra, J. J., Liebrand, W. B. G., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour and Information Technology, 17*(3), 155–63.

[6] Fogg, B. J & Tseng, H(1999). The elements of computer credibility. In Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, CHI '99, pages 80–87, New York, NY, USA, 1999. ACM.

[7] Fogg, B. J. (2003). Persuasive technology : using computers to change what we think and do. Amsterdam ; Boston: Morgan Kaufmann Publishers, /.

[8] Gretzel, U., & Fesenmaier, D. (2006). Persuasion in Recommender Systems. Int. J. Electron. Commerce, 11(2), 81-100.

[9] Herlocker, J.; Konstan, J.; and Riedl, J. (2000). Explaining collaborative filtering recommendations. In W.A. Kellogg and S. Whittaker (eds.), *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work.* Philadelphia: ACM Press,pp. 241–250.

[10] Hovland, C. I., & Weiss, W. The influence of source credibility on communication effectiveness. *Public Opinion Quarterly,* 1951, *15,* 635-650.

[11] Jimison, H. B., Street, R. L., Jr., & Gold, W. R. (1997). Patient-specific interfaces to health and decision-making information, *LEA's communication series*. Mahwah, NJ: Lawrence Erlbaum.

[12] Norman, D.A. (2004) *Emotional Design: Why We Love(or Hate) Everyday Things.* New York: Basic Books.

[13] Nowak, G. J., Shamp, S., Hollander, B., Cameron, G. T., Schumann, D. W., & Thorson, E. (1999). Interactive media: A means for more meaningful advertising? *Advertising and consumer psychology*. Mahwah: Lawrence Erlbaum.

[14] Pu, P., & Chen, L. (2007). Trust-inspiring explanation interfaces for recommender systems.*Knowledge-BASED Sytems, 20*, 542-556.

[15] Pu, P., & Chen, L. (2010). A user-centric evaluation framework of recommender systems. Proceedings of the ACM RecSys 2010 Workshop on User-Centric Evaluation of Recommender System and Their Interfaces(UCERSTI), Barcelona, Spain, Sep 30.

[16] Shneiderma, B. (2003). Promoting Universal Usability with Multi-Layer Interface Design. *Proceedings of the 2003 Conference on Universal Usability.*ACM New York

[17] Spiekermann, S., and Paraschiv, C.(2002). Motivating human-agent interaction: Transferring insights from behavioral marketing to interface design. *Journal of Electronic Commerce Research, 2*, 255–285.

[18] Strecher, V. J. (1999). Computer-tailored smoking cessation materials: A review and discussion.Special Issue: Computer-tailored education. *Patient Education & Counseling, 36*(2), 107– 117.

[19] Strecher, V. J., Kreuter, M., Den Boer, D.-J., Kobrin, S., Hospers, H. J., & Skinner, C. S. (1994). The effects of computer-tailored smoking cessation messages in family practice settings. *Journal of Family Practice, 39*(3), 262–270.

[20] Swearingen, K.& Sinha, R.(2001) Beyond algorithms: An HCI perspective on recommender systems. In *Proceedings of the SIGIR 2001 Workshop on Recommender Systems*.